語末の形態的特徴に基づく日本語派生語対の収集*

加藤 直樹† 藤田 篤‡ 佐藤 理史‡
†名古屋大学工学部電気電子・情報工学科 ‡名古屋大学大学院工学研究科
naoki@sslab.nuee.nagoya-u.ac.jp, {fujita,ssato}@nuee.nagoya-u.ac.jp

1 はじめに

言語には同じ意味を表す複数の表現が存在する．したがって，計算機で言語を柔軟に扱うには，異なる表現が同じ意味を持っていると認識したり，一方から他方を自動生成したりする言い換え技術は不可欠である．

様々な言い換え現象が，社会的な背景とは独立に，言語に関する知識だけを用いて説明できる可能性がある [5]．頑健な言い換え処理のために必要な，言語知識のうち，語に関する知識，語と語の関係に関する知識（語彙知識）は重要な役割をになう．例えば，品詞は異なるが関連する意味を持つ語対に関する知識があれば，例 (1) に示すような言い換えを扱うことができる．

(1) a. 部屋は十分**暖かい**⇔ 部屋は十分**暖まっている**
b. 身体**がだるいと**感じる ⇔ 身体**のだるさを**感じる
c. 円**の**レート**が下がった**⇔ 円**が**レート**を下げた**

語間の関係を記述する枠組の一つに，Meaning-Text Theory [11] がある．これは，語間の関係を語彙関数と呼ばれる形式で記述し，それらを組み合わせることで言い換えや機械翻訳を実現する理論である．ただし，語間のあらゆる関係を記述するのにどれだけの種類の語彙関数が必要なのか，各言語において語彙関数の実体となる語対をいかに収集するか，などの問いに対する解は，経験的に明らかにするしかない．

英語に関しては，語間の関係を扱う大規模な資源が構築されてきた．例えば，WordNet [3] や CatVar [4] は，同義，対義，派生，上位–下位，部分–全体，論理的含意などの関係を扱っており，上のような言い換えの処理にも利用できる．一方，日本語に関しては，語の同義および上位下位関係を収録したシソーラスは存在するが，それ以外の関係に関する十分に大規模な資源は，我々が知る限り存在しない．

このような背景から，我々は，派生語対，すなわち「暖かい」–「暖まる」,「楽しい」–「楽しみ」のように，語幹を共有し，明確に意味的関連のある語対を，正確かつ大規模に収集する手法を検討している．これまでに，語末に形態的特徴を持つ派生語対を既存の単語辞書およびコーパスから収集し，表 1 に示すような仕様の派生語辞書を構築した．本稿では，派生語対の収集手法および構築した辞書の仕様について述べる．

表 1: 派生語辞書のエントリ例

語 1	品詞 1	語 2	品詞 2	語幹	派生パターン	派生元	表記語幹
暖かい	*A*	暖まる	*V*	atata	＊-kai:＊-maru	語 1	暖
：	：	：	：	：	：	：	
楽しい	*A*	楽しみ	*N*	tanosi	＊-i:＊-mi	語 1	楽し
：	：	：	：	：	：	：	

2 派生関係を表す手がかり

Habash ら [4] は，英語における派生語辞書 CatVar を，

1. 人間用の単語辞書，形態素解析用の辞書などの 6 種類の既存の辞書から，大規模な語のリストと小規模な派生語対を得る．
2. 既知の派生語対および Porter stemmer で共通の語幹を持つ複数の語を一つのクラスタにまとめる．

という手順で構築した．この手法にならい，我々も，まず形態素解析器を用いて彼らと同じ手法で派生語対を獲得することを思い浮かべた．しかし，形態素解析用の辞書が形態素と語の両方を含んでいる [14] ことから，形態素解析器を stemmer として利用することは困難と考え，形態素解析用の辞書を広義の語の集合とみなすに留めることにした．そして，まずは形態素，語，派生語の扱いに関する知見を語形成に関する文献に求めた．

斎藤 [13] は，形容詞「高い」を例に,「高」を語幹（文献中は語基）とする派生語の形態的特徴を派生プロセスと呼んでまとめている．また，伊藤ら [6] は,「高い」から「高まる」を作る過程を,『接辞付加（affixation）による語形成』としている．以上より下記の知見を得た．

- 接辞ごとに，付加できる語の品詞が決まっている．
- 接辞という形態的特徴によって派生語を作るプロセスが説明できる．
- 「＊-i:＊-maru」のような派生を表す形態的特徴は，「高い」–「高まる」や「広い」–「広まる」といった複数の派生語対に共通であるが，あらゆる語に対して適用可能ではない．

これらをふまえ，派生語間の語末に見られる「＊-i:＊-maru」のような形態的特徴（以下，**派生パターン**と呼ぶ）を派生語対の収集に用いることにした．

*Collecting Derivatives based on Suffix Patterns of Word Pairs.
Naoki Kato†, Atsushi Fujita‡, Satoshi Sato‡
†Department of Electrical and Electronic Engineering and Information Engineering, School of Engineering, Nagoya University
‡Graduate School of Engineering, Nagoya University

3　派生語対候補収集

予想される派生語対の規模や形態的特徴の多様性を考慮し，品詞対ごとに，まず派生パターンを，続けて派生語対の候補を収集する．

3.1　派生パターンの収集

まず，次の 3 ステップで派生パターンを収集する．

Step 1. 2 つの品詞の単語辞書から語幹を共有する異品詞の語対を網羅的に取り出す．単語辞書として与えるべき情報は，表記とローマ字表記の 2 つである．ここで，語幹を共有するとは，語に含まれる漢字，およびローマ字表記の先頭 1 文字以上が完全に一致することを指す．例えば，「押し付ける」–「押し付けがましい」は語幹「押付，ositsuke」を共有する．

Step 2. 各語対について，ローマ字表記の差分を結合したものを派生パターンとする．例えば，上の語対からは「＊-ru:＊-gamasii」という派生パターンを得る．

Step 3. 各派生パターンを、対応する語対の異なり数（サポート数）とともに出力する．

3.2　派生語対候補の収集

次に，2 つの品詞の単語辞書（仕様は 3.1 項の Step 1 に同じ）から，個々の派生パターンに合致する語対を，次の 3 ステップで網羅的に取り出す．

Step 1. 語幹を共有し，語幹以外の部分が 3.1 項で得た派生パターンのいずれかに一致する語対を抽出する．

Step 2. 仮名の送り方が等しい語対のみを抽出する．

抽出条件 1. 語対の表記の差分をローマ字に変換したものと，派生パターンが一致する．

抽出条件 2. 語対の表記の差分をローマ字に変換したものと，派生パターンにパターン直前のローマ字一文字を加えたものが一致する．

例えば，「懐く」–「懐しい」の表記の差分「く」–「しい」のローマ字表記「ku」–「sii」は，派生パターン「＊-u:＊-asii」とは一致せず（条件 1），派生パターンに直前のローマ字「k」を加えたパターン「＊-ku:＊-kasii」とも一致しない（条件 2）ので棄却する．

Step 3. 語幹が一致し，語全体の読みが一致するものは異表記としてまとめて出力する．

3.3　抽出結果

内容語（名詞，動詞，形容詞，形容動詞，副詞）の各品詞対を対象として，形態素解析器用辞書 IPADIC（MeCab 版）[1] から，上記の手順で派生語対を収集した．得られた全ての派生パターンの集合 P，それに対応する候補語対の集合 D，2 つ以上の語対に対応する派生パターンの集合 P_1，およびそれに対応する候補語対の集合 D_1 の規模を表 2 に示す．パターンの数よりも語対の数の方が少ない場合があるが，これは，派生語対候補を収集する際に，送り仮名が一致していることを条件に加えたためである．

表 2: 辞書から得られた派生語候補対の規模

ID 品詞対	P	D	P_1	D_1
(a) 名詞–動詞	6,190	7,902	486	3,275
(b) 名詞–形容詞	1,600	1,676	74	299
(c) 名詞–形容動詞	1,020	2,331	34	1,447
(d) 名詞–副詞	784	839	32	252
(e) 動詞–形容詞	994	979	72	244
(f) 動詞–形容動詞	534	459	28	72
(g) 動詞–副詞	652	632	26	71
(h) 形容詞–形容動詞	286	287	18	63
(i) 形容詞–副詞	137	138	9	30
(j) 形容動詞–副詞	113	145	4	65
合計	12,310	15,388	783	5,818

名詞–動詞対は動詞とその連用形の対であり大規模に得られた．名詞–形容動詞対は 1,363 対が同形の語対であったが，IPADIC のマニュアル [1] では，これらは意味が違うとされている．名詞–副詞対は，「今日」のような『名詞–副詞可能』を除いてあったが，「案の定」，「多少」のように名詞と副詞の両方の用法を持つ同形の語対が 109 語得られた．その他の語対の内容については，50 語対以上に対応するパターンはなかった．

4　被覆についての考察

既存の資源に対する被覆について考察し，被覆の向上のために，コーパスを用いて 2 形態素以上からなる語の派生語対を収集した．

4.1　IPAL 日本語単語辞書との比較

計算機用日本語辞書 IPAL [7, 8]（以下 IPAL）は，個々の語（正確には語義）の振る舞いの分析を目的として，いくつかの言語テストの結果を収録した電子的資料である．各語の形態に関する情報として，比較的緩やかな基準で派生語が収集されている．今回は，IPAL 形容詞辞書 [8] に掲載されている形容詞と各品詞の語対，および IPAL 動詞辞書 [7] の動詞と各品詞の語対の和集合を I とし，我々の，異表記をまとめる前の語対候補，すなわち 3.2 項の Step 2 の出力の集合 D' との共通部分および差分を調査した．結果を表 3 に示す．

被覆は，30.2%と，予想していたほどは高くならなかった．これは，2 節で述べたように，我々が入力とした形態素辞書と IPAL におけるエントリの単位の違いにある．たとえば，「＊-sa:＊-i」という派生パターンに合致する名詞–形容詞対は，IPAL には 172 語対含まれていたが，形態素辞書の名詞辞書には，わずか 10 語しか掲載されていなかった．

[1] http://mecab.sourceforge.jp/, ver. 2.7.0-20060707

表 3: IPAL の派生語表記対に対する被覆

ID 品詞対	$\lvert D'\rvert$	$\lvert D'_I\rvert$	$\lvert D' \cap I\rvert$	$\lvert I_{D'}\rvert$	$\lvert I\rvert$
(a) 名詞–動詞	8,078	7,666	412	367	779
(b) 名詞–形容詞	1,694	1,619	75	287	362
(c) 名詞–形容動詞	2,467	2,460	7	49	56
(d) 名詞–副詞	859	859	0	0	0
(e) 動詞–形容詞	987	869	118	331	449
(f) 動詞–形容動詞	463	458	5	37	42
(g) 動詞–副詞	636	614	22	68	90
(h) 形容詞–形容動詞	290	286	4	313	317
(i) 形容詞–副詞	141	139	2	41	43
(j) 形容動詞–副詞	150	144	6	12	18
合計	15,765	15,114	651	1,505	2,156

$D'_I = D' \setminus I, I_{D'} = I \setminus D'$, 被覆 $= |D' \cap I| \,/\, |I| = 0.302$

表 4: コーパスからの候補収集に用いる派生パターン

派生の向き	派生パターン (辞書から得られた語対の数)
形容詞→名詞	＊-i:＊-mi (36)， ＊-i:＊-me (18)， ＊-i:＊-sa (10)
名詞→形容詞	＊-ϕ:＊-ppoi (12)， ＊-ϕ:＊-rasii (5)
形容詞→形容動詞	＊-i:＊-ge (8)， ＊-i:＊-sou (0)

4.2 コーパスを用いた派生語対候補収集

2 形態素からなる語を含む語対を，Langkilde ら [10] の考え方を参考にして作成した

典型的な派生パターンで生成された表現が実際に良く使われるのであれば，正しい語であり，元の語の派生語である．

という仮説に基づいて収集した．今回は，表 4 に示す派生パターンを用いて “派生語らしい” 表現を自動生成し，各々について毎日新聞データ集 1991〜2005 年版における出現回数を調べた．この際，生成した表現が当該品詞として振る舞うことを保証するための条件を加えた．例えば，名詞を生成した場合は格助詞が後続することを，形容動詞を生成する場合は「だ」,「で」,「な」のいずれかが後続することを条件とした．

コーパスにおける出現回数が 2 以上の語と元の語の対の集合を C，また $O = D' \cup C$ とする．表 5 に示すように，2 形態素以上からなる語を扱うことで当該品詞における被覆の度合を向上させることができた．被覆率をさらに向上させる工夫は後で述べるようにいくつか考えられるが，まず，派生パターンに基づく手法で収集できる語対の範囲を調査した．具体的には，提案手法に照らして，一部の品詞対の $I \setminus O\,(= I_O)$ に含まれる語対を次のように分類した（表 6 も参照のこと）．

- S_1 **対象外の語対：** 漢字以外で始まる語を含む語対，語頭の接辞による派生語対，語幹を共有してない語対．
- S_2 **現在の方法では収集不可能な語対：** 「寒い」–「寒ざむと」のように繰り返しによる派生語対．
- S_3 **既知の派生パターンと合致する語対：** 派生パターンが P または表 4 に含まれる語対．
- S_4 **新しい派生パターンに合致する語対：** P と表 4 のどちらにも含まれていない派生パターンを持つ語対．

表 5: コーパスから得られた派生語対候補を含めた場合の被覆

ID 品詞対	$\lvert O\rvert$	$\lvert O_I\rvert$	$\lvert O \cap I\rvert$	$\lvert I_O\rvert$	$\lvert I\rvert$
(b) 名詞–形容詞	3,431	3,203	228	134	362
(h) 形容詞–形容動詞	468	361	107	210	317
小計	3,899	3,564	335	344	679
合計	17,680	16,773	907	1,249	2,156

$O_I = O \setminus I, I_O = I \setminus O$, 被覆 $= |O \cap I| \,/\, |I| = 0.420$

表 6: 未収集の派生語対の細分類（括弧内はパターン数）

ID 品詞対	S_1	S_2	S_3	S_4	$\lvert I_O\rvert$
(b) 名詞–形容詞	53	0	75 (3)	6 (4)	134
(e) 動詞–形容詞	65	5	90 (18)	171 (8)	331
(f) 動詞–形容動詞	12	0	3 (2)	22 (2)	37
(g) 動詞–副詞	4	32	26 (9)	6 (5)	68
(h) 形容詞–形容動詞	78	4	128 (3)	0 (0)	210
(i) 形容詞–副詞	5	26	9 (2)	1 (1)	41
合計	217	67	331 (37)	206 (20)	821

S_2 に含まれる語対はすべて語幹部分の繰り返しであり，パターンの抽象化の必要性が明らかになった．

S_3 に含まれる語対の多くは，語対候補の収集源とした単語辞書に語が含まれていなかったため得られなかった．(b)，(h) のうちコーパスから獲得できなかった 62 件，97 件についても，単にコーパスの規模の問題ではなく，派生語候補の生成元となる語が辞書に存在しなかった可能性がある．これらをふまえ，今後は語対候補の収集源として，より大きな単語辞書を用いることを検討する．例えば，形態素解析器 JUMAN 用の辞書[2] は，長単位の形態素を扱っており，IPADIC に収録されていない「四角い」–「四角ばる」などを含んでいる．表記の揺れの網羅性を高めるために，EDR 日本語単語辞書 [12] を用いることも考えられる．

ただし，辞書の大規模化にしたがってパターン数が増えた場合は，パターンそのものの吟味が必須である．例えば，(e)，(f) については,「＊-sugiru:＊-i」,「＊-sugiru:＊-ϕ」というパターンに合致する語対が，S_4 にそれぞれ 161 件，21 件含まれていたが，これらは派生接辞による派生語というよりも動詞「すぎる」との複合語である．我々は，語対の一方のみが複合語となる場合は派生語対とは考えないので，このパターンは不要である．

5 精度についての考察

派生パターンを用いるだけでは,「見つける」–「見にくい」のように明確な意味的関連のない語対も収集してしまう．そこで，派生語対を選別するために，いくつかの設問集合からなる図 1 の判定基準を設けた．

表 2 の D_1 の派生語対候補のうち，表 6 の 6 品詞対について，作業者 2 人が完全に独立に，図 1 の設問にしたがって良否のラベルを付与した．判定対象 779 語対のうち，2 人ともが派生語対である (A) と判定したのは 428 件，2 人ともが派生語対でない (C) と判定した

[2] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

設問 1. 各語の意味が分かる．
　Yes: 設問 3 へ．No: 設問 2 へ．
設問 2. 少なくとも一方が語ではない．
　Yes: 派生語対でない（C1）．No: 判定保留（B1）．
設問 3. 語幹の表記・読みともに共通である．
　Yes: 設問 4 へ．No: 派生語対でない（C2）．
設問 4. 各語の品詞は正しい．
　Yes: 設問 5 へ．No: 派生語対でない（C3）．
設問 5. 一方のみが，複合語または可能・使役の接辞を持つ．
　Yes: 派生語対でない（C4）．No: 設問 6 へ．
設問 6. 語間に明確な意味的な関連性がある．
　Yes: 設問 7 へ．No: 派生語でない（C5）．
設問 7. 派生元はどちらか．
　左: A1，右: A2，分からない: A3，
　両方が同じ語からの派生語: A4．

図 1: 派生語対の判定基準

のは 147 件，1 人以上が意味がわからない (B) と判定したのは 59 件，A と C で意見が分かれたのは 145 件であり，1 人以上が派生語対であると判断したものは約 74%にとどまった．これは，Habash ら [4] が報告している 92%という精度を大きく下回る結果となった．

精度を改善するには，C に分類される語対を自動的に排除する必要がある．C1，C3 に分類される語対は，語対候補の収集源を，解析の頑健性を重視した形態素解析用辞書から，より規範的な語彙を収録した辞書に変えることで排除できると期待できる．また，C4 に分類される語対も，単独の語と完全に一致する語対を含む派生パターンを用いないことにすれば排除できる．しかし，意味的な関連性の有無（2 名とも C5 とした語対は 74 件）を機械的に判定することは難しく，最終的な辞書編纂には人手での良否の判定は必須である．

1 人以上が B と判定した候補語対を除く 720 件について，2 人の被験者による判定の一致率 P_o，およびその偶然性を考慮した κ 値 [2] を算出した（表 7）．表中，P_{o_9} および κ_9 はラベルの細分類が一致している度合，P_{o_2} および κ_2 は A または C という判定が一致している度合を指す．κ_9 および κ_2 は 2 人の判定の一致率が中程度であったことを示している．主に，語と非語，品詞の適否，および単一の語か複合語か，の各設問において作業者間の判定の違いが生じていたため，今後，図 1 の判定基準について再検討する予定である．

表 7: 良否判定の一致率

ID 品詞対	A/C の語対数	細分類の一致 P_{o_9}	κ_9	A/C の一致 P_{o_2}	κ_2
(b) 名詞–形容詞	269	0.59	0.48	0.79	0.55
(e) 動詞–形容詞	235	0.80	0.73	0.85	0.65
(f) 動詞–形容動詞	68	0.25	0.10	0.63	0.10
(g) 動詞–副詞	67	0.78	0.56	0.81	0.48
(h) 形容詞–形容動詞	54	0.54	0.36	0.74	0.35
(i) 形容詞–副詞	27	0.96	0.94	0.96	0.87
全体	720	0.66	0.55	0.80	0.53

6　おわりに

本稿では，形態的特徴に基づいて派生語対候補を収集する手法を提案した．語末に着目した派生パターンを整理し，それに合致する語対を収集した．また，良否判定の基準を作り，一部の語対候補の良否を判定した．

今後は，規模の拡大と並行して収録情報を豊かにすることも検討する．「＊-i:＊-garu」というパターンで表される「広い」–「広がる」と「怖い」–「怖がる」は意味的には異なる種類の派生であると考えられる．このような派生の意味を辞書に注釈付けし，例 (1) のような言い換えの生成・認識の実現を目指す [9].

本研究の一部は次の研究費の支援を受けている：科研費基盤研究 (A)「円滑な情報伝達を支援する言語規格と言語変換技術」（課題番号：16200009，代表：佐藤理史）および科研費若手研究 (B)「文法カテゴリ交替を裏付ける語彙特性の体系化と辞書記述」（課題番号：18700143，代表：藤田篤）．

参考文献

[1] 浅原正幸, 松本裕治. ipadic version 2.7.0 ユーザーズマニュアル. Technical report, 奈良先端科学技術大学院大学, 2003.

[2] R. Bakeman and J. M. Gottoman. *Observing interaction: an introduction to sequential analysis (second edition)*. Cambridge University Press, 1997.

[3] C. Fellbaum. A semantic network of English verbs. In C. Fellbaum, editor, *WordNet: an electronic lexical database*. The MIT Press, 1998.

[4] N. Habash and B. J. Dorr. A categorial variation database for English. In *Proceedings of the 2003 Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics* (*HLT-NAACL*), pp. 17–23, 2003.

[5] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理, Vol. 11, No. 5, pp. 151–198, 2004.

[6] 伊藤たかね（編）. 文法理論:レキシコンと統語. 東京大学出版会, 2002.

[7] 情報処理振興事業協会技術センター. 計算機用日本語動詞辞書 IPAL (Basic Verbs) －解説編－. 情報処理振興事業協会技術センター, 1987.

[8] 情報処理振興事業協会技術センター. 計算機用日本語形容詞辞書 IPAL (Basic Adjectives) －解説編－. 情報処理振興事業協会技術センター, 1990.

[9] 加藤修平, 藤田篤, 佐藤理史. 句を対象とした構成的な言い換えの生成. 2007. (in this proceedings).

[10] I. Langkilde and K. Knight. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics* (*COLING-ACL*), pp. 704–710, 1998.

[11] I. Mel'čuk. Lexical functions: a tool for the description of lexical relations in a lexicon. In L. Wanner, editor, *Lexical Functions in Lexicography and Natural Language Processing*, pp. 37–102. John Benjamin Publishing Company, 1996.

[12] 日本電子化辞書研究所. EDR 電子化辞書仕様説明書. 1995.

[13] 斎藤倫明. 現代日本語の語構成論的研究. ひつじ書房, 1992.

[14] 佐藤理史. 境界認定の提案: (1) コンセプトと実現方法. 情報処理学会研究報告, NL-164-5, pp. 25–32, 200.